QUARTERLY REPORT

to the

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Contract NSR 39-011-076
Quarterly report #7

University of Pittsburgh
Knowledge Availability Systems Center

July 1969

INTRODUCTION

The "Reporting Requirements" for Contract No. NSR 39-011-076, "Contract for special experimental projects involving information systems and technology utilization," specify that each of the six projects involved are to be accounted for separately in reporting. However, in this, as in recent quarterly reports, only four of the six projects under this contract will be reported. In response to a request from the National Aeronautics and Space Administration (NASA), the Knowledge Availability Systems Center (KASC) of the University of Pittsburgh has completed the work and furnished a final report for projects #4, Small Business Administration Participants Study, and #6, Study of Fee-Paying Industrial Client Attrition.* The four projects covered by this report are:

Project #1: Investigation of Means for Improving Exploitation of Indexes

Project #2: Thesaural Development to Permit Improved Relevance Predictability of Searches

Project #3: Investigation of Comparative Relevance of Search Results

Project #5: Demand Searches

Project #1

Investigation of Means for Improving Exploitation of Indexes

I. Background

As has been previously reported, the initial development of this study that has as its objective the "...consolidation of similar searches in order to achieve production economies and the improvement of confidence

*. Knowledge Availability Systems Center, "Attrition" in Information Dissemination Relationships with Industry, National Aeronautics and Space Administration, Contract NSR 39-011-076, Final Report, 1968.

in exhaustiveness of searches for industry over and above what may be achieved through searches of the indexes as provided by the government," the need for a natural text searching technique was recognized.

Since Project #3, Investigation of Comparative Relevance of Search Results, also has need of a technique for searching natural text, progress on this project is reported only once (page 4).

Project #2

Thesaural Development to Permit Improved Relevance Predictability of Searches

I. Background

The objective of this study is to determine the applicability of the National Aeronautics and Space Administration Thesaurus to the file of documents indexed prior to the implementation of the Thesaurus,* i.e.; during the period the Subject Authority List/(SAL) was in effect.

II. Procedures

Two aspects of applicability have been considered, (1) the relationships between the forms of the terms found in the Thesaurus and the forms found in the SAL, and (2) the effect of the structure of the Thesaurus on retrieval when the subterms listed under a "Main Term", which is also a strategy term, are substituted for a term in a strategy originally developed for searching the retrospective file.

In order to assure that the term substitutions were made consistently throughout the experiment, a subset of the Thesaurus of particular applicability to the National Aeronautics and Space Administration/Regional Dissemination Center (NASA/RDC) at the University of Pittsburgh was created by converting the terms in the original SAL based strategies to the form of the term found in the Thesaurus, and developing a SAL/Thesaurus consisting of the "Main Terms", which were also strategy terms, in the Thesaurus and all of the appropriate subterms. The terms

* January 1968.

in this subset of the Thesaurus were then converted to the form of the term found in the SAL.

In addition to the original strategies whose terms were used to develop the subset of the Thesaurus, the Knowledge Availability Systems (KAS) Center has available as a data base (1) the document citations that resulted from the use of the strategies in the computer search of the file during one sample "current awareness" search period* and, (2) two levels of evaluation for the cited documents, the analysts' and the users' evaluations of the documents forwarded following analysts' review.

In order to determine the retrieval effectiveness of the subterms listed under the original strategy terms, single aspect searches of the sample period were made for the subterms and the result of substituting the subterm for the original term in the strategy was recorded. A comparison was made of the effectiveness of the original strategy and the strategy modified by the substitution of a subterm for the original strategy term.

III. Analysis and Results

An analysis of the results of the comparisons achieved by the substitution of the subterm for the original term in the strategy is now in process.

In addition, the subset of the Thesaurus, the SAL/Thesaurus, in which the terms in the form found in the SAL appearing in the display used in the Thesaurus, is being corrected prior to its final listing.

A doctoral dissertation relating to this aspect of the effort has been completed, and defended successfully, and copies are now being prepared for submission as a final report.

---

*Knowledge Availability Systems Center, University of Pittsburgh, "Special Experimental Projects Involving Information Systems and Technology Utilization," Quarterly Progress Report to the National Aeronautics and Space Administration, No. 4 on Contract No. 39-011-076, Oct 68.

Project #3

Investigation of Comparative Relevance of Search Results

I. Summary of Earlier Reports*

The efficiency and effectiveness of retrieving documents by means of index terms has been questioned for some time. The utility of index terms, in general, has been questioned, as well as, the reliability of literature searches conducted by mechanized systems using index terms. It is this last question that prompted this investigation. Traditionally, professional indexers have assigned descripters to documents, and, in the course of performing their professional function, have acted as an interperator of the meaning of the author's words to the user. It is possible that through this process the concepts of the author are corrupted by (1) the use of different terminology to describe identical or partially related concepts; (2) use of the same term to describe different but realted concepts; (3) the varying depth of indexing which may lead to uneven penetration of the subject content of documents; and, (4) the varying subject knowledge of the indexers which leads to inconsistencies in analysis of documents. It is felt that these factors may effect the discriminating and recall capabilities of a system.

An experiment has been designed**and conducted to empirically test the hypothesis that the text of certain, well defined, portions of documents in the NASA file (hereafter called document surrogates) can be searched either singly or in combinations with the same effectiveness as would result from searching the file on index terms alone. Previously formulated search strategies were used, and current procedures for mechanized searches of the file were employed. The search strategies were specifically constructed for the searching of the NASA file by means of index terms.

---

*Knowledge Availability Systems Center, University of Pittsburgh, "Special Experimental Projects Involving Information Systems and Technology Utilization," Quarterly Progress Reports to the National Aeronautics and Space Administration, Nos. 1, 2, 3, 4 and 5 on Contatct No. 39-011-076. Jan 68, Apr 68, Jly 68, Oct 68 and Jan 69.

**Description of the experimental design has been reported in quarterly reports 1-5, ibid.

II. The Experiment

A. Design

1. File: The document surrogates* of 1,195 NASA and IAA generated documents were keypunched into machine-readable form to create the experimental file. A program for an IBM 360/40 computer was used which deleted common words (functional words and articles), and then inverted the experimental file, and rearranged it alphabetically to form a concordance. The concordance contains the non-deleted text terms as well as coded information for each text term which indicates which document surrogate, sentence in the document surrogate, and word in the sentence it represents.

   An illustration of the concorded file was presented in Quarterly Report No. 5.**

2. Search Strategies:

   From the population of over 600 search strategies a sub-population of 256 strategies which met defined criteria (discussed in previous reports***) was identified. A sample of 50 strategies was randomly selected from the sub-population, and these were used to search the experimental file.

B. Procedures

To conduct the investigation of the concordance, three reviews were deemed necessary. The initial review required an exact match between the text word and the search strategy term. For the second review the plural ending of the term (s or es) was masked, resulting in a deviation from the exact match with respect to number. It was assumed

*Title, abstract, first paragraph, last paragraph, notation of content, and index terms.

**Knowledge Availability Systems Center, University of Pittsburgh, "Special Experimental Projects Involving Information Systems and Technology Utilization," Quarterly Progress Report to the National Aeronautics and Space Administration, No 5, on Contract No. 39-011-076, Jan 69.

***Reported in Reports 1-5, ibid.

that the plural of a term had the same meaning as the singular form of the term except for number. The third review deviated even farther from the exact match. Here variant endings of terms were disregarded and the text term and strategy term were considered as a match if the stems matched. Such endings as ed, ing, al, ally, etc. are included in the group of variant endings.

The modifications of strategy terms were identified to determine how much deviation from an exact match can be tolerated and still maintain the required precision of meaning.

Additional information has been collected for each of these reviews. For example, the contiguity relationships derived from the location of the text term in the document surrogate can be determined, and, if they indicate something of interest, they can be analyzed in detail.

Anticipated Outcome

It is expected that by searching the text of document surrogates rather than index entries, the recall capabilities of the system will be increased. If the increase is only minimal it will offer at least the same recall value obtained by searching the index entries for these same documents. It is also anticipated that the discriminating capabilities of the system will be increased by searching text terms, but that the level of discrimination will decrease rapidly as the modification of the strategy term increases, thereby increasing the degree of mismatch between the strategy term and the text term.

Project #5

Demand Searches

During the reporting period, the National Aeronautics and Space Administration (NASA) requested the Knowledge Availability Systems (KAS) Center to perform retrospective searches in the NASA unclassified document collection for 35 profiles.

The searches were performed using computer index tapes provided by NASA to the KAS Center in its regional dissemination activities, and the results were transmitted to the following individuals:

	Number of Profiles
Mr. F. R. Hand IIT Research Institute 10 West 35th Street Chicago, Illinois 60616	33
Mr. Roy Bivins, Jr. Office of Technology Utilization National Aeronautics and Space Administration Washington, D. C. 20546	2

Three types of service were provided for the 35 profiles:

Type I - The addressee received a copy of the printout of the computer search listing only the accession numbers of the documents identified.

Type II - The addressee received a copy of the printout of the computer search listing only the accession numbers of the documents identified, plus a copy of the abstract of each document or, in the case of Tech Briefs, a copy of the cited Brief.

Type III - The addressee received a listing of the accession numbers of the documents related to the profile, cited mechanically or manually, plus copies of the abstracts of the related documents, or, in the case of Tech Briefs, a copy of the relevant Brief.

Of the 33 profiles delivered to Mr. F. R. Hand, 24 were provided with Type I service and 9 were provided with Type II service. The two profiles delivered to Mr. Roy Bivins, Jr. both recieved Type III service.

The strategy terms used in performing the searches were provided both by the recipients of the search results and by KAS Center subject specialists. Mr. Hand provided terms based on the NASA Thesaurus for the 33 profiles which he received. The terms were appropriate for the 1968-1969 portion of the file, but required conversion by KAS Center personnel to the NASA Subject Authority List for the 1962-1967 portion of the file. For the two profiles delivered to Mr. Bivins, strategy terms were selected by the KAS Center staff based on the descriptions of the profiles received from Mr. Bivins and, for one profile, from a Dr. David Foster.

A Boolean expression of the selected terms for each profile was keypunched by KAS Center staff and submitted with the KAS Center search program and the appropriate computer tape to the University of Pittsburgh Computation Center for execution of the search using its IBM 7090 system. The structure of the KAS Center computer tape file requires that a complete retrospective search for a profile be performed in three computer runs.

Although the profiles were batched, the different times of submission of the 35 profiles required that a total of 9 computer runs be performed.

The searches for all 35 profiles resulted in the citing of 3,993 documents of which 3,961 were cited mechanically and 32 were identified manually. The sources of the documents, by service type, was as follows:

SOURCE	TYPE I	%	TYPE II	%	TYPE III	%	TOTAL	%
IAA	1,295	41.9	202	51.9	230	44.8	1,727	43.2
Aerospace Medicine	169	5.5	1	.3	16	3.1	186	4.7
STAR	1,623	52.5	185	47.5	264	51.5	2,072	51.9
Tech Briefs	4	.1	1	.3	3	.6	8	.2
TOTALS	3,091	100.0	389	100.0	513	100.0	3,993	100.0

The 3,091 documents identified in the Type I service were in response to the strategies used for 22 profiles. Strategies used for 2 of the 24 profiles which were provided with this service resulted in the message "No citations". copy of the computer printout for each profile was forwarded to Mr. Hand.

The 389 documents identified in the Type II service were in response to the strategies used for 8 profiles. The strategy selected for one of the 9 profiles receiving this service resulted in the message "No citations".

Copies of the abstracts of the documents cited from IAA, Aerospace Medicine, and STAR were produced and forwarded to Mr. Hand along with a copy of the single Tech Brief and a copy of the computer printout for each of the 9 profiles in the Type II service.

Of the 513 documents identified in Type III service, 481 were identified mechanically and 32 were identified manually. Follwoing the computer search performed for the two profiles receiving this service, copies of the abstracts of the identified documents were produced and delivered, with the computer printout and a description of the profile, to KAS Center subject specialists who weeded irrelevant documents from the output. Three hundred (58.5%) documents were eliminated by this process.

Using the cumulative subject indexes of IAA and STAR, the specialists manually identified 13 relevant documents from IAA and 19 relevant documents from STAR which had not been cited mechanically. These were added to the 213 documents remaining from the computer searches all of which were then forwarded with a listing of the documents to Mr. Bivins. The distribution of the forwarded documents by source was as follows:

IAA	109
Aerospace Medicine	4
STAR	129
Tech Briefs	3
TOTAL	245

No evaluation of the searches by the recipients has been received by the KAS Center other than a comment by Mr. Bivins who stated that he understood Mr. Hand to be highly pleased with the results that he received.